ISSN No:-2456-2165

A Machine Learning Approach to Improve the Cement Manufacturing Process by Optimising the Time for Quality Checking

Ashwini KS[1], Mihir Jain[2], Ankita Yadav[3], G.M V N Pavan Kumar[4], Bharani Kumar Depuru[5]

[1]Research Associate, Innodatatics, Hyderabad, India.
[2]Research Associate, Innodatatics, Hyderabad, India.
[3]Mentor, Research and Development, Innodatatics, Hyderabad, India.
[4]Team Leader, Research and Development, Innodatatics, Hyderabad, India
[5]Director, Innodatatics, Hyderabad, India
***Corresponding Author:** Bharani Kumar Depuru
ORC ID: 0009-0003-4338-8914

Abstract:- The cement manufacturing comprises a series of steps aimed at producing top-tier cement that adheres to industry benchmarks while minimising residual content. Traditional practices involve periodic quality assessments, often hourly, facilitated by sensor-derived data.

Cement quality assessment hinges on two critical parameters: residue and reject, which gauge cement fineness. Residue reflects the non-uniformity in the final cement output, influenced by various input factors. Thus, meticulous tracking of pertinent inputs is essential. Yet, this method's drawback is that substandard cement mandates the rejection of entire batches. Even with sensor automation, this approach remains time-intensive, less accurate and detrimental to productivity, culminating in substantial losses encompassing raw materials, time, labour, revenue, and in-demand market fulfilment.

To surmount these challenges, the integration of automation in quality assessment with machine learning processes, buoyed by adept algorithms, has emerged as an efficient solution. The pivotal target lies in condensing the quality check time frame from an hour to a mere minute, thereby necessitating computational intelligence. Machine learning models offer a path to automate quality checks, dramatically curtailing the time investment compared to conventional methods. Leveraging historical data from companies, these models are trained to streamline the process.

In this context, the deployment of a regression model proves invaluable for predicting and anticipating cement residue, a dependable gauge of its quality. Training the regression model with extensive datasets confers it with the power to discern residue and reject levels accurately, classifying cement quality through the analysis of diverse factors including raw material composition, production parameters, and environmental conditions. The model can unveil hidden patterns and correlations that influence residue levels. This empowers manufacturers to rapidly evaluate the quality of cement batches and expedite corrective actions when necessary.

The project employs diverse datasets to train various regression models including multi-linear regression, K-nearest neighbour regression, decision trees, random forest, adaboost, xgboost, and neural network models like multi-layer perceptron. The next step involves evaluating the efficiency and accuracy of these trained models, with a focus on selecting relevant metrics. Given the objective of forecasting residue and reject levels, the mean absolute percentage error (MAPE) is adopted. A lower MAPE value indicates more precise predictions, with a targeted MAPE value set below 10%. To address high MAPE values for the "reject" variable, ensemble stacking of models is employed, involving meticulous hyperparameter tuning for each algorithm. This stacking amalgamates predictions from multiple models, yielding enhanced accuracy.

Integrating K-nearest neighbour regression, random forest, and xgboost algorithms in a stacked ensemble capitalises on individual model strengths while offsetting their limitations, ultimately refining the residue and reject level predictions. By implementing the aforementioned strategies, the production process is optimised, quality control practices are heightened, and waste is minimised. The result is the delivery of superior-quality cement that seamlessly aligns with market demand.

Keywords:- *Cement Manufacturing, Cement Quality Management, Regression Model, Machine Learning, Time Optimization, Quality-Checking.*

ISSN No:-2456-2165

I. INTRODUCTION

The objective of this research study is to enhance the efficiency of the cement manufacturing process through the implementation of a machine learning (ML) model. The industry currently grapples with issues related to delays in quality checking, leading to significant losses in materials, time, labor, and revenue. These challenges directly impact the ability to meet market demands. The primary focus of this research project is to address and resolve the delay issue in quality checking. Traditionally, the quality checking of cement involved the use of the sieve method. After the clinker was manufactured and cement extracted as output, it underwent a sieving process. The residue, representing non-cement particles resulting from manufacturing errors, was collected, and the percentage in the cement was recorded along with reject values. Even with sensor automation, this process proved to be time-consuming, requiring approximately an hour for each quality checking cycle. Furthermore, quality checks could only be conducted after the completion of the manufacturing process. To overcome these challenges and optimise the time involved, regression models were employed. This research study and development follow the Cross-Industry Standard Process for Machine Learning with Quality Assurance available on the open-source framework of 360DigiTMG website [Fig. 1][1].

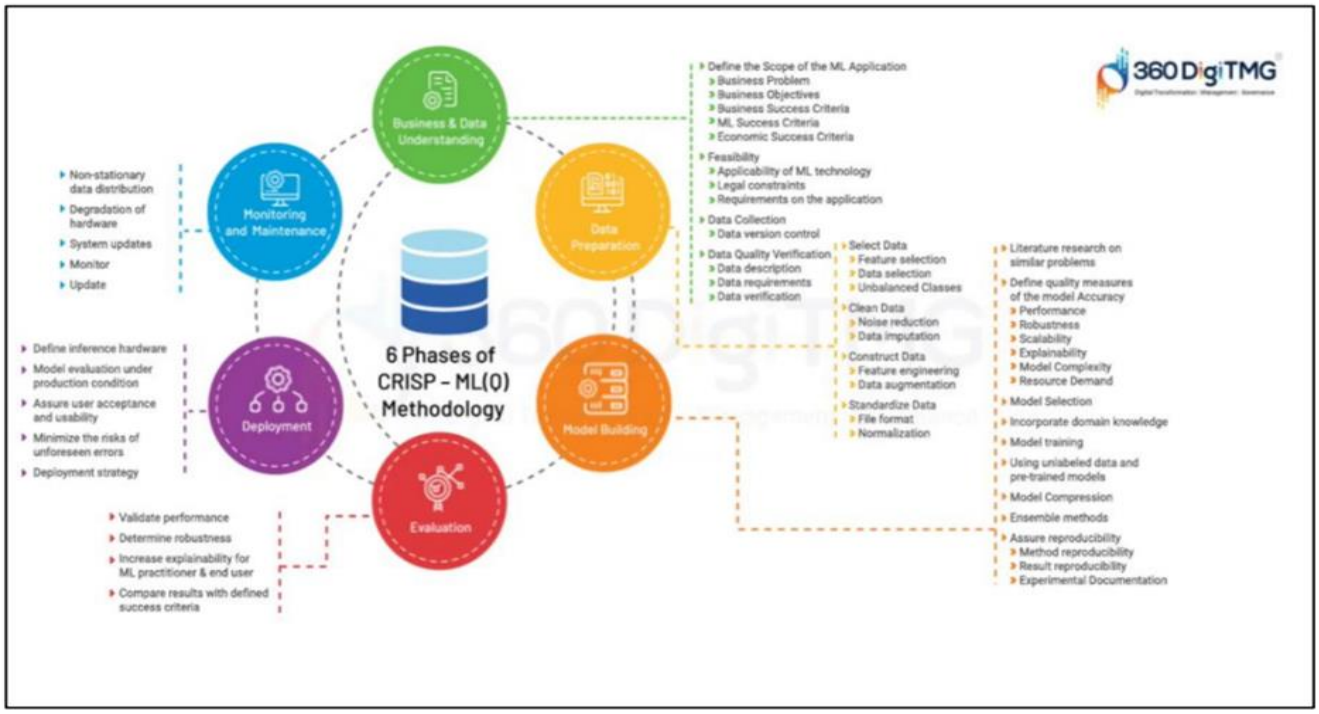

Fig.1: CRISP-ML(Q) : Cross-Industry Standard Process for Machine Learning and Quality Assurance (Source: CRISP-ML(Q) | 360DigiTMG)

By incorporating automation, decision-making processes undergo simplification through computational methodologies such as machine learning models [2]. These approaches enable computers to acquire knowledge and make predictions without explicit programming, leading to a transformative shift in various industries. Regression models [3], a subset of diverse machine learning algorithms, are specifically tailored for predicting numerical data output.

The random forest [3] emerges as a generative bagging ensemble technique, utilising multiple decision trees to produce accurate outcomes. In contrast, XGBoost, a boosting ensemble technique, optimises predictions by utilising residuals and relies on a series of decision trees [4] during training. The Multilayer Perceptron (MLP), a foundational artificial neural network [5] [6] architecture in machine learning and deep learning, demonstrates proficiency in solving complex problems [7].

Stacking is an ensemble technique [7] in machine learning, enhances predictive performance by combining forecasts from diverse base models [Fig.2]. The process involves:

- Base Models: Training diverse base models using different algorithms or subsets of data.
- Predictions: Generating forecasts on the same dataset using these base models.
- Meta-Model: Training a meta-model (often a simple linear regression or decision tree) based on the forecasts made by base models.
- Final Prediction: The meta-model integrates predictions from base models to make the final prediction, typically achieving higher accuracy and generalisation compared to individual models.

Stacking leverages the strengths of diverse models, amplifying overall performance and serving as a valuable tool in machine learning.

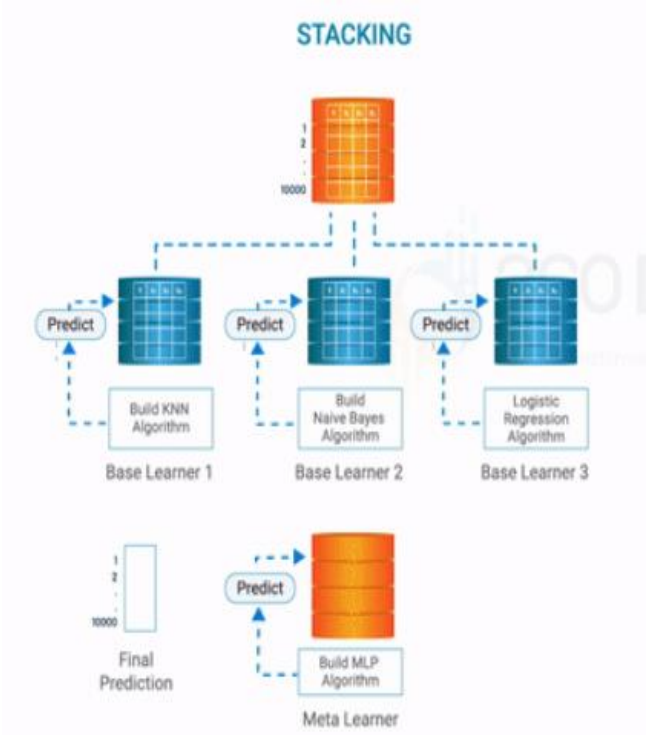

Fig.2: *Stacking Ensemble (non-generative model) : Base learner and Meta learner stacked.*
(Source: https://360digitmg.com/animated-learning)

II. METHODS AND TECHNIQUES:

The whole research project was performed using a standardised flow that is followed in the industry as shown in the figure [Fig.3].

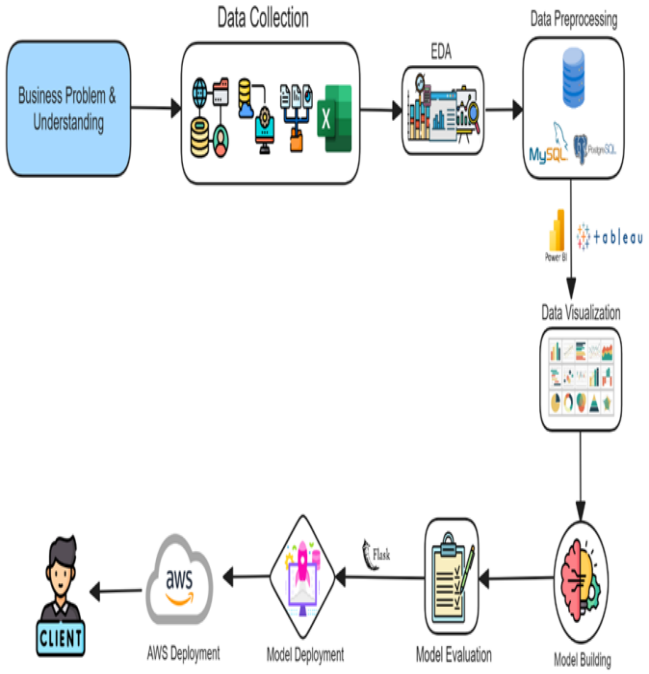

Fig.3*: Project Architecture - Flow and stages of project*

➢ *Data Collection*

The journey begins with data collection. In the cement manufacturing process [8], this involves gathering a wealth of information on various parameters. These parameters include raw material composition, production processes, energy consumption, temperature, humidity, and the target variables residue and reject. By collecting comprehensive data, manufacturers gain insights into the complex interplay of variables that affect the quality and efficiency of their processes.

Client data majorly consists of two distinct folders named Process Data and QC Data - SPOT. Each excel file within consists of date-wise sheets. The Process Data folder consists of excel sheets that contain input parameters recorded based on date and time, representing the inputs provided to the cement manufacturing process. QC DATA-SPOT folder contains the data within these monthly files including date and residue columns, where the residue column represents the quality of cement produced. Our primary task was to merge the data from these folders based on the common date and time parameters.

➢ *Data Mapping*

The next major step is to map data to convert it into a usable structure. The dataset for this project was obtained from two distinct folders provided by the client. This process involves creating a structured representation of the collected information. Data mapping not only organises the data but also establishes relationships between different variables, aiding in more profound analysis [9]. It serves as a crucial bridge between raw data and actionable insights. Data mapping consists of matching and merging residue value from one file to date and time value from another.

➢ *Data Preprocessing*

Data preprocessing is the critical step of cleaning, transforming, and organising the data to ensure its reliability and usability. This stage may involve identifying and removing outliers [10], duplicates, filling in missing values [10], standardising data formats and feature selection or extraction. These manipulations are carried out using various python libraries.

Duplicates were dropped, missing values were handled using *median imputation* to maintain data completeness without altering central tendencies. Negative entries were identified and replaced with their *absolute values*. Outliers were managed through winsorization techniques. Additionally, *min-max* scaling was applied to standardise feature values, promoting uniformity.

➢ *Exploratory Data Analysis And Visualisation:*

Exploring the data set for its various statistical features gives us a picture on the overall nature of this data. It can be easily achieved using libraries in python. Auto-visualisation can be done for examining the same using Auto-EDA libraries. Further, checking for various irregularities in data and fixing them accordingly is carried out. Data types, distinct variables, histogram, central tendencies, skewness, kurtosis, box-plots, qq-plots, correlation heatmap etc was

ISSN No:-2456-2165

performed on each column before and after the preprocessing to check quality of data as it is essential for accurate analysis and model development. [11]

Next crucial step in making data understandable, no matter how clean and structured, can be overwhelming in its raw form. Visualisation comes to the rescue by translating data into easily understandable charts, graphs, and dashboards. These visual representations allow manufacturers to spot trends, anomalies, and correlations at a glance. For cement manufacturers, this means having real-time insights into key quality indicators [4], [12].

Predicting and optimising with refined data and clear visualisations, the process moves to model building.

➢ *Model Building*

Analytical and predictive models are created during this stage. Here, machine learning algorithms are crucial to help manufacturers streamline their workflows, cut down on manufacturing time and error, and maintain or even raise quality [13], [14].

The target variable is residue and reject which are necessary variables for cement quality checking [15], which are both continuous or numerical types of data. Hence, the regression model, with models being stacked using ensemble technique [16], was trained using the above collected data.

To conduct training of data, the dataset was divided into training data and testing data at the ratio of 7:3. The next crucial step is to obtain the model that gives the best accuracy. We trained several models using a training dataset. Linear regression models, KNN, Support Vector Regressor, Decision tree regressor, Ensemble techniques were trained. Further, hyperparameter tuning is carried out using hyperopt to obtain the best set of hyperparameters. To test for accuracy, several error matrices were considered. Looking at the behaviour of the dataset, Mean Absolute Percentage Error (MAPE) [12] value for the error matrix was finally considered.

The successful creation of predictive models paves the way for their deployment into the manufacturing process. This integration is where theory meets practice. The models developed in the previous stages are put to work, guiding real-time decisions on raw material mixing, production parameters, and quality checks. This step is critical in achieving the dual goals of time optimization and quality control. By tracking key performance indicators and comparing them against the predictions made by the models, manufacturers can make necessary adjustments and fine-tune their processes. This ongoing feedback loop ensures the sustainability of time optimization efforts.

➢ *Hyperparameter Tuning And Model Comparison:*

- **Random Forest:** 10000 trials, run for 43 h, 09 min, 40 s
  Hyperparameters obtained after tuning for residue as output are;
  'Criterion': ‘squared_error, 'max_depth': 13, 'max_features': 2, 'min_samples_leaf': 9.0, 'min_samples_split': 16.0, 'n_estimators': 100.0, 'oob_score': True

- **XGBoost:** 10000 trials, run for 2 h, 17 mins, 19 s
  Hyperparameters obtained after tuning for residue as output are;
  'alpha': 5.0, 'colsample_bylevel': 0.2686726185810717, 'colsample_bytree': 0.46114459629796245, 'gamma':0.6307597103680819, 'lambda': 4.0, 'learning_rate': 0.047555089081800464, 'max_depth':10, 'n_estimators': 120.0, 'subsample': 0.8617910846379223

- **KNN Regressor:** 10000 trials, run for 2 h, 04 mins, 24 s
  Hyperparameters obtained after tuning for residue as output are;
  'leaf_size': 55.0, 'metric': 1, 'n_neighbors':27.0, 'weights': ‘uniform

- **MLP Regressor:** 10000 trials, run for 11 h, 09 mins, 43 s
  Hyperparameters obtained after tuning for residue as output are;
  'activation': ‘relu’, 'alpha': 6.300000000000001, 'batch_size': auto, 'early_stopping': False, 'hidden_layer_sizes': 2, 'learning_rate': constant, 'solver': 'lbfgs'

III. RESULTS AND DISCUSSION:

For the results, we would first explore the data and discuss the observations made. The monthly pattern of average residue was first observed. In this section, we would majorly observe the trends and models keeping residue as the output [Fig.4]. The received data contains information of two years, 2022 and 2023.

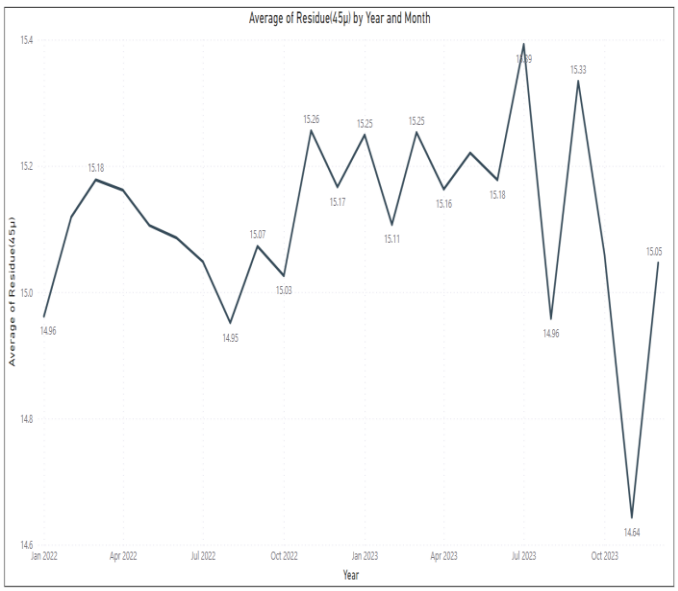

Fig.4: *Monthly pattern of average residue, range of residue is 12-14%*

ISSN No:-2456-2165

The output is well balanced as residue shows normal distribution. It was confirmed by the following histogram plot [Fig.5] and density plot [Fig.6].

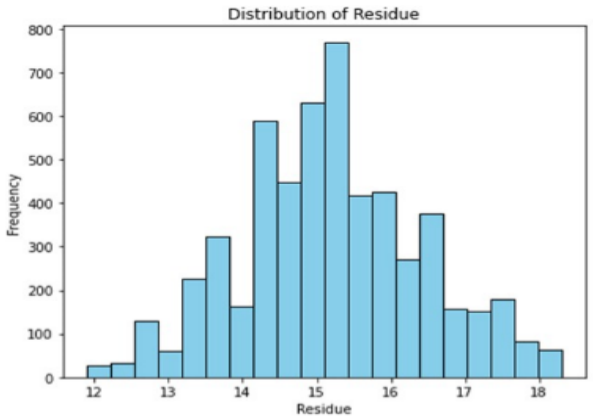

***Fig.5**: Histogram of the output residue to infer on data distribution*

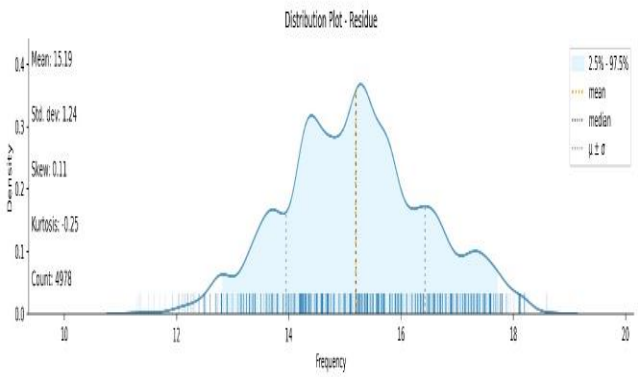

***Fig.6**: Density Plot of residue revealing the data to undergo normal distribution*

Correlation matrix and heatmap for residue is produced. Using the correlation heatmap, as shown in [Fig.7], we decided to choose the top features. A few features were chosen based on domain knowledge as well. The resulting data set was then split into training and testing data at the ratio of 7:3 respectively.

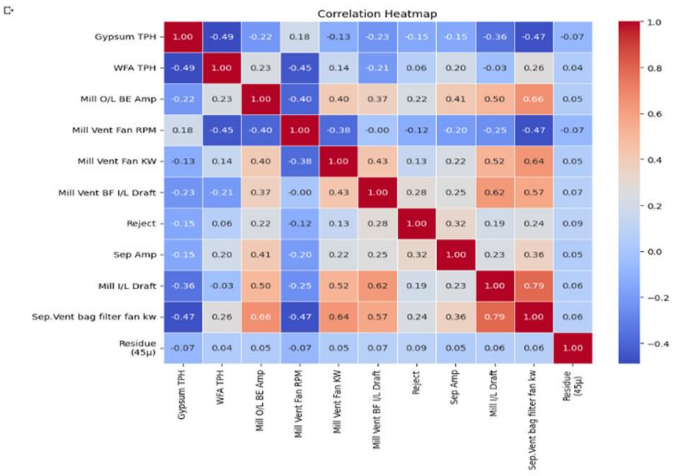

***Fig.7**: Correlation Heatmap that shows the correlation of input variables with the output Residue*

After various trials, based on the mean absolute percentage error (MAPE), the best model was chosen. Decision tree and ensemble techniques showed practical and acceptable results. The table below [Fig.8]shows MAPE values for each of the models. As we may observe, when stacking ensemble technique was used, with base learners as K-nearest neighbour regressor, random forest and XGBoost, and meta learner as multi-layer perceptron was used, we observed the best results with least error and least variance. Hence, the best model was chosen.

Sr.No	MODEL	TRAIN MAPE	TEST MAPE
1	KNN Regressor	6.212134805	6.210861688
2	XGBoost	5.354697919	6.092875998
3	Random Forest	4.60977026	6.066264452
4	MLP Regressor	6.234012076	6.277644197
5	**STACKING ENSEMBLE(MLP Regressor as Meta Model)**	**4.438104648**	**4.335939609**

***Fig.8**: Mean Absolute Percentage Errors for models with residue as output.*

The stacking ensemble model with meta learner as multi-layer perceptron regressor was run to get the best hyperparameters using hyperopt, with 10,000 trials, for 11 hours, 9 mins, and 43s. However, when deployed using AWS globally or even on Flask locally, the results are obtained in just around 1 min, meeting the criteria. Thereby, using this model, the client can predict the residue values and hence, realise the quality and reduce errors in quality checking.

The major concern in a project is the accuracy and error. From the above table in [Fig.8], we may infer that the MAPE value of 4.43% of training accuracy for residue was obtained, meaning an accuracy of 95.57% is attained. Also, the correlation between actual and predicted residue using this model reaches a strong 0.70 score. These percentages meet our machine learning success criteria and hence making the development a success. The outcomes would directly result in cost-saving, exceeding $1 million, which is a significant impact.

IV. CONCLUSION

In this research study, a problem with quality checking in a cement manufacturing industry was identified and dealt with. The amalgamation of automated quality assessment with machine learning processes offers a groundbreaking solution to the challenges faced by traditional cement manufacturing methods. The conventional practice of hourly quality assessments, relying on sensor-derived data, proves time-intensive and results in substantial losses in terms of raw materials, time, labour, revenue, and meeting market

demands due to the rejection of entire batches. The proposed solution involves harnessing machine learning models, particularly regression models, to predict crucial parameters such as cement residue and reject levels. These models are trained with comprehensive datasets encompassing diverse factors like raw material composition, production parameters, and environmental conditions, enabling manufacturers to expedite the assessment of cement quality and take prompt corrective actions.

The project explores a range of regression models, showcasing a thorough examination of computational intelligence. The use of the mean absolute percentage error (MAPE) as a metric for model evaluation, with MAPE value 4.43% attained, ensures precision in predicting the residue levels. The innovative approach of ensemble stacking further enhances accuracy by combining predictions from multiple models, effectively addressing the limitations of individual algorithms.

The optimization of the production process, heightened quality control practices, and minimise waste collectively result in the delivery of superior-quality cement that aligns seamlessly with market demand. This successful integration of machine learning and manufacturing processes stands as evidence of the transformative potential of advanced technologies in revolutionising and optimising traditional industrial practices.

REFERENCES

[1]. S. Studer *et al.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, Apr. 2021, doi: 10.3390/make3020020.

[2]. M. Bhandari and B. Silwal, "Development of Machine Learning Model Applied to Industrial Motors for Predictive Maintenance," in *2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*, Bengaluru, India: IEEE, Nov. 2022, pp. 1632–1635. doi: 10.1109/IIHC55949.2022.10060358.

[3]. S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

[4]. A. K. Udugu and D. A. Khare, "Automation of Cement Industries," vol. 1, no. 6, 2014.

[5]. S. Banihashemi, S. Khalili, M. Sheikhkhoshkar, and A. Fazeli, "Machine learning-integrated 5D BIM informatics: building materials costs data classification and prototype development," *Innov. Infrastruct. Solut.*, vol. 7, no. 3, p. 215, Jun. 2022, doi: 10.1007/s41062-022-00822-y.

[6]. I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. in Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2016.

[7]. C.-H. Chen, K. Tanaka, M. Kotera, and K. Funatsu, "Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications," *J. Cheminformatics*, vol. 12, no. 1, p. 19, Dec. 2020, doi: 10.1186/s13321-020-0417-9.

[8]. E. H. Gautier, I. R. Hurlbut, and E. A. E. Rich, "Recent Developments in Automation of Cement Plants," *IEEE Trans. Ind. Gen. Appl.*, vol. IGA-7, no. 4, pp. 458–469, Jul. 1971, doi: 10.1109/TIGA.1971.4181327.

[9]. "The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf."

[10]. N. Kumar, V. Naranje, and S. Salunkhe, "Cement strength prediction using cloud-based machine learning techniques," *J. Struct. Integr. Maint.*, vol. 5, no. 4, pp. 244–251, Oct. 2020, doi: 10.1080/24705314.2020.1783122.

[11]. P. Bruce, A. Bruce, and P. Gedeck, "Practical Statistics for Data Scientists".

[12]. M. Mohtasham Moein *et al.*, "Predictive models for concrete properties using machine learning and deep learning approaches: A review," *J. Build. Eng.*, vol. 63, p. 105444, Jan. 2023, doi: 10.1016/j.jobe.2022.105444.

[13]. W. Z. Taffese and E. Sistonen, "Machine learning for durability and service-life assessment of reinforced concrete structures: Recent advances and future directions," *Autom. Constr.*, vol. 77, pp. 1–14, May 2017, doi: 10.1016/j.autcon.2017.01.016.

[14]. B. Hökfors, M. Eriksson, and E. Viggh, "Modelling the cement process and cement clinker quality," *Adv. Cem. Res.*, vol. 26, no. 6, pp. 311–318, Dec. 2014, doi: 10.1680/adcr.13.00050.

[15]. A. K. Mishra and A. Jha, "Quality Assessment of Sarbottam Cement of Nepal".

[16]. [S. Tibshirani and H. Friedman, "The Elements of Statistical Learning Data Mining, Inference, and Prediction.".